IDS 702: Module 1.2

Introduction to multiple linear regression

Dr. Olanrewaju Michael Akande

Multiple linear regression

- Multiple linear regression (MLR) assumes the following distribution for a response variable y_i given p potential covariates/predictors/features $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n.$$

- We can also write the model as:

$$y_i \overset{iid}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}, \sigma^2).$$

$$p(y_i|\boldsymbol{x}_i) = \mathcal{N}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}, \sigma^2).$$

- MLR assumes that the conditional average or expected value of a response variable is a linear function of potential predictors.

- Note that the linearity is in terms of the "unknown" parameters (intercept and slopes).

- Just like in SLR, MLR also assumes values of the response variable follow a normal curve within any combination of predictors.

MLR

- Just as we had under SLR, here each β_j represents the true "unknown" value of the parameter, while $\hat{\beta}_j$ represents the estimate of β_j.

- Similarly, y_i represents the true value of the response variable, while $\hat{y}_i$ represents the predicted value. That is,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta} x_{i2} + \ldots + \hat{\beta} x_{ip}.$$

- Also, the residuals e_i are our estimates of the true "unobserved" errors ϵ_i. Thus,

$$e_i = y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta} x_{i2} + \ldots + \hat{\beta} x_{ip}\right] = y_i - \hat{y}_i.$$

- Since the e_i's estimate the ϵ_i's, we expect them to also be independent, centered at zero, and have constant variance.

- We will get into this more under model assessment.

IDS 702

MLR: ESTIMATION

- Estimated coefficients are found by taking partial derivatives of the sum of squares of the errors

$$\sum_{i=1}^{n} \left(y_i - [\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}]\right)^2,$$

  with respect to each parameter, that is, $\beta_0, \beta_1, \ldots, \beta_p$.

- This is the ordinary least squares (OLS) method.
- Resulting formulas are a bit messy to write down in this form.
- However, there is a very nice matrix algebra representation as we will see soon.

MLR: ESTIMATION

- An alternative derivation uses maximum likelihood estimation (MLE).
- First, not that if each Y_i, with $i = 1, \ldots, n$, follows the normal distribution $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, then the likelihood is

$$L(\mu, \sigma^2 | y_1, \ldots, y_n) = \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(y_i - \mu)^2}$$

$$= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2}.$$

- So that for MLR, the likelihood is

$$L(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2 | y_1, \ldots, y_n) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - [\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}])^2}.$$

- To get the MLEs, take the log of the likelihood, differentiate with respect to each parameter in $(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2)$, and set to zero.
- Again, resulting formulas for $(\beta_0, \beta_1, \ldots, \beta_p)$ are a bit messy to write down in this form.

IDS 702

MLR: ESTIMATION

- The MLE for σ^2 (work it out to convince yourself) is

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}\right]\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2.$$

- However, the MLE is biased. That is, $\mathbb{E}[\hat{\sigma}^2_{\text{MLE}}] \neq \sigma^2$.
- Therefore, we often used the following "unbiased" estimator for σ^2.

$$\hat{\sigma}^2 = s_e^2 = \frac{1}{n-(p+1)}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-(p+1)}\sum_{i=1}^{n} e_i^2.$$

- Most software packages will estimate s_e^2 automatically.

IDS 702

MLR: MATRIX REPRESENTATION

- Let

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad \boldsymbol{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

- Then, we can write the MLR model as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \;\; \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}).$$

- The OLS and MLE estimates of all $(p+1)$ coefficients (intercept plus p slopes) is then given by

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

Ideally, n should be bigger than p. Why?

There are many ways around the $p > n$ problem. If there is time, we may look at some options.

IDS 702

MLR: MATRIX REPRESENTATION

- The predictions can then be written as

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left[\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}\right] = \left[\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right]\boldsymbol{y}.$$

- The residuals can be written as

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \left[\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right]\boldsymbol{y} = \left[\mathbf{1}_n - \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right]\boldsymbol{y}$$

  where $\mathbf{1}_n$ is a matrix of ones

- The $n \times n$ matrix

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$$

  is often called the projection matrix or the hat matrix.

- We will see some important features of the elements of $\boldsymbol{H}$ soon.

IDS 702

MLR: MATRIX REPRESENTATION

- In matrix form,

$$s_e^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n-(p+1)} = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-(p+1)} = \frac{\boldsymbol{e}^T\boldsymbol{e}}{n-(p+1)}.$$

- The variance of the OLS estimates of all $(p+1)$ coefficients (intercept plus p slopes) is

$$\mathbb{V}\left[\hat{\boldsymbol{\beta}}\right] = \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$$

- Notice that this is a covariance matrix; the square root of the diagonal elements give us the standard errors for each β_j, which we can use for hypothesis testing and interval estimation.

What are the off-diagonal elements?

- When estimating $\mathbb{V}[\hat{\boldsymbol{\beta}}]$, plug in s_e^2 as an estimate of σ^2.

- Now that we have a basic introduction, we are ready see how to fit MLR models.

IDS 702

What's next?

Move on to the readings for the next module!